Repeated polyploidization of *Gossypium* genomes and the evolution of spinnable cotton fibres

Andrew H. Paterson[1], Jonathan F. Wendel[2], Heidrun Gundlach[3], Hui Guo[1], Jerry Jenkins[4,5], Dianchuan Jin[6], Danny Llewellyn[7], Kurtis C. Showmaker[8], Shengqiang Shu[4], Joshua Udall[9], Mi-jeong Yoo[2], Robert Byers[9], Wei Chen[6], Adi Doron-Faigenboim[10], Mary V. Duke[11], Lei Gong[2], Jane Grimwood[4,5], Corrinne Grover[2], Kara Grupp[2], Guanjing Hu[2], Tae-ho Lee[1], Jingping Li[1], Lifeng Lin[1], Tao Liu[6], Barry S. Marler[1], Justin T. Page[9], Alison W. Roberts[12], Elisson Romanel[13], William S. Sanders[8], Emmanuel Szadkowski[2], Xu Tan[1], Haibao Tang[1,14], Chunming Xu[2,15], Jinpeng Wang[6], Zining Wang[1], Dong Zhang[1], Lan Zhang[6], Hamid Ashrafi[16], Frank Bedon[7], John E. Bowers[1], Curt L. Brubaker[7,17], Peng W. Chee[18], Sayan Das[1], Alan R. Gingle[1], Candace H. Haigler[19], David Harker[9], Lucia V. Hoffmann[20], Ran Hovav[10], Donald C. Jones[21], Cornelia Lemke[1], Shahid Mansoor[1,22], Mehboob ur Rahman[22], Lisa N. Rainville[1], Aditi Rambani[9], Umesh K. Reddy[23], Jun-kang Rong[1], Yehoshua Saranga[24], Brian E. Scheffler[11], Jodi A. Scheffler[11], David M. Stelly[25], Barbara A. Triplett[26], Allen Van Deynze[16], Maite F. S. Vaslin[27], Vijay N. Waghmare[28], Sally A. Walford[7], Robert J. Wright[29], Essam A. Zaki[30], Tianzhen Zhang[31], Elizabeth S. Dennis[7], Klaus F. X. Mayer[3], Daniel G. Peterson[8], Daniel S. Rokhsar[4], Xiyin Wang[1,6] & Jeremy Schmutz[4,5]

Polyploidy often confers emergent properties, such as the higher fibre productivity and quality of tetraploid cottons than diploid cottons bred for the same environments[1]. Here we show that an abrupt five- to sixfold ploidy increase approximately 60 million years (Myr) ago, and allopolyploidy reuniting divergent *Gossypium* genomes approximately 1–2 Myr ago[2], conferred about 30–36-fold duplication of ancestral angiosperm (flowering plant) genes in elite cottons (*Gossypium hirsutum* and *Gossypium barbadense*), genetic complexity equalled only by *Brassica*[3] among sequenced angiosperms. Nascent fibre evolution, before allopolyploidy, is elucidated by comparison of spinnable-fibred *Gossypium herbaceum* A and non-spinnable *Gossypium longicalyx* F genomes to one another and the outgroup D genome of non-spinnable *Gossypium raimondii*. The sequence of a *G. hirsutum* A_tD_t (in which 't' indicates tetraploid) cultivar reveals many non-reciprocal DNA exchanges between subgenomes that may have contributed to phenotypic innovation and/or other emergent properties such as ecological adaptation by polyploids. Most DNA-level novelty in *G. hirsutum* recombines alleles from the D-genome progenitor native to its New World habitat and the Old World A-genome progenitor in which spinnable fibre evolved. Coordinated expression changes in proximal groups of functionally distinct genes, including a nuclear mitochondrial DNA block, may account for clusters of cotton-fibre quantitative trait loci affecting diverse traits. Opportunities abound for dissecting emergent properties of other polyploids, particularly angiosperms, by comparison to diploid progenitors and outgroups.

The *Gossypium* genus is ideal for investigating emergent consequences of polyploidy. A-genome diploids native to Africa and Mexican D-genome diploids diverged ~5–10 Myr ago[4]. They were reunited ~1–2 Myr ago by trans-oceanic dispersal of a maternal A-genome propagule resembling *G. herbaceum* to the New World[2], hybridization with a native D-genome species resembling *G. raimondii*, and chromosome doubling (Fig. 1). The nascent A_tD_t allopolyploid spread throughout the American tropics and subtropics, diverging into at least five species; two of these species (*G. hirsutum* and *G. barbadense*) were independently domesticated to spawn one of the world's largest industries (textiles) and become a major oilseed.

New insight into *Gossypium* biology is offered by a genome sequence of *G. raimondii* Ulbr. (chromosome number, 13) with ~8× longer scaffold N50 (18.8 versus 2.3 megabases (Mb)) compared with a draft[5], and oriented to 98.3% (versus 52.4%[5]) of the genome (Supplementary Table 1.3a). Across 13 pseudomolecules totalling 737.8 Mb, ~350 Mb (47%) of euchromatin span a gene-rich 2,059 centimorgan (cM), and ~390 Mb (53%) of heterochromatin span a repeat-rich 186 cM (Supplementary Discussion, sections 1.5 and 2.1). Despite having the least-repetitive DNA of the eight *Gossypium* genome types, *G. raimondii* is 61% transposable-element-derived (Supplementary Table 2.1). Long-terminal-repeat retrotransposons (LTRs) account for 53% of *G. raimondii*, but only 3% of LTR base pairs derive from 2,345 full-length elements. The 37,505 genes and 77,267 protein-coding transcripts annotated (Supplementary Table 2.3 and http://www.phytozome.com) comprise 44.9 Mb (6%) of the genome, largely in distal chromosomal regions (Supplementary Discussion, section 2.1).

Shortly after its divergence from an ancestor shared with *Theobroma cacao* at least 60 Myr ago[6], the cotton lineage experienced an abrupt five- to sixfold ploidy increase. Individual grape chromosome segments resembling ancestral eudicot genome structure, or corresponding cacao chromosome segments, generally have five
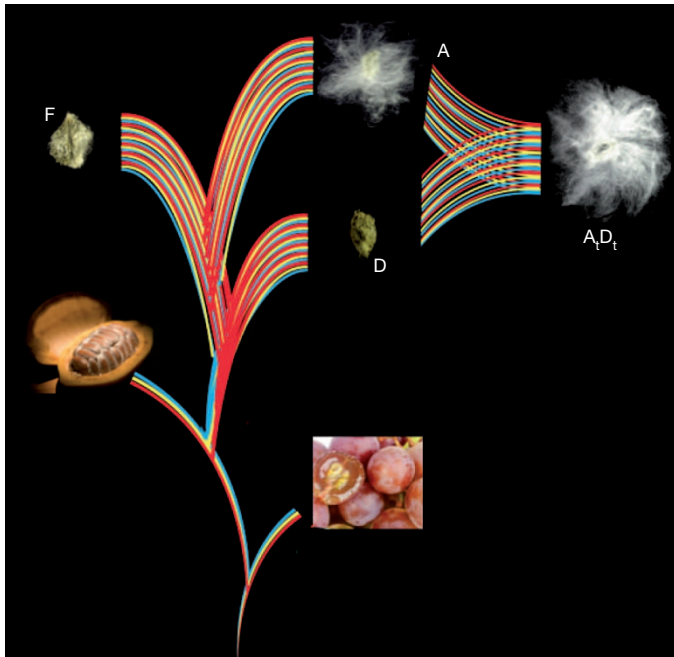
Figure 1 | Evolution of spinnable cotton fibres. Paleohexaploidy in a eudicot ancestor (red, yellow and blue lines) formed a genome resembling that of grape (bottom right). Shortly after divergence from cacao (bottom left), the *Gossypium* lineage experienced a five- to sixfold ploidy increase. Spinnable fibre evolved in the A genome after its divergence from the F genome, and was further elaborated after the merger of A and D genomes ~1–2 Myr ago, forming the common ancestor of *G. hirsutum* (Upland) and *G. barbadense* (Egyptian, Sea Island and Pima) cottons.

ploidy increase show a single peak of synonymous nucleotide-substitution (K_s) values, suggesting either one, or multiple closely spaced, event(s) (Supplementary Fig. 3.5). Pairwise cytological similarity among A-genome chromosomes[9] suggests the most recent event was a duplication.

Paleopolyploidy may have accelerated cotton mutation rates: for 7,021 co-linearity-supported gene triplets, K_s rates and non-synonymous nucleotide-substitution (K_a) rates were, respectively, 19% and 15% larger for cotton–grape than cacao–grape comparisons (Supplementary Table 3.2). Adjusted for this acceleration (Supplementary Fig. 3.5), the cotton ploidy increase occurred about halfway between the pan-eudicot hexaploidy (<125 Myr ago[10]) and the present, near the low end of an estimated range of 57–70 Myr ago[11].

Paleopolyploidy increased the complexity of a Malvaceae-specific clade of Myb family transcription factors, perhaps contributing to the differentiation of epidermal cells into fibres rather than the mucilages of other Malvaceae. Among 204 R2R3, 8 R1R2R3 and 194 heterogeneous Myb transcription factors in *G. raimondii* (Supplementary Table 3.5), subgroup 9 has six members known only in Malvaceae (Fig. 3a), comprising a possible 'fibre clade' distinct from the *Arabidopsis thaliana* GL1-like subgroup 15 involved in trichome and root hair initiation and development[12]. Expressed predominantly in early fibre development, elite cultivated tetraploid cottons have higher expression of five (50%) of ten subgroup 9 genes compared with wild (undomesticated) tetraploids (Fig. 3a and Supplementary Table 5.3). Some subgroup 9 genes are also active in leaves, hypocotyls and cotyledons (Supplementary Fig. 3.8), consistent with specialization for different types of epidermal cell differentiation such as production of a 'pulpy layer' secreted from the teguments surrounding cacao seeds, and mucilages in other Malvaceae fruit (*Abelmoschus* (okra), *Cola* (kola)) and roots (*Althaea* (marshmallow)).

Cotton growers were early adopters of integrated pest management[13] strategies to deploy intrinsic defences conferred by pest- and disease-resistance genes that evolved largely after the 5–6-fold ploidy increase.

(infrequently six) best-matching *G. raimondii* regions and secondary matches resulting from pan-eudicot hexaploidy[7,8] (Fig. 2 and Supplementary Table 3.1). Paralogous genes tracing to this five- to sixfold

Figure 2 | Syntenic relationships among grape, cacao and cotton. a, Macrosynteny connecting blocks of >30 genes (grey lines). Highlighted regions (pink and red) trace to a common ancestor before the pan-eudicot hexaploidy[7], with the *Gossypium* lineage five- to sixfold ploidy increase forming multiple derived regions. Inferred duplication depth in cotton varies (top). **b,** Micro-synteny of grape chromosome (Chr) 3, cacao chromosome 2 and five cotton chromosomes. Rectangles represent predicted genes, with connecting grey lines showing co-linear relationships. An example (1 grape, 1 cocoa, 5 cotton) is highlighted in red.

Figure 3 | Paleo-evolution of cotton gene families. a, Myb subgroup 9 (ref. 12) originated from a gene on the progenitor of cacao chromosome 2 that formed two adjacent copies after Malvales–Brassicales divergence and then triplicated in cotton, with subsequent loss of one chromosome 8 and two chromosome 12 paralogues. One extant paralogue traces to pan-eudicot hexaploidy, Tc04 g009420, and reduplicated in cotton (Gorai.012G052500.1 and Gorai.011G122800.1) and *Arabidopsis*[8] (At3g01140 and At5g15310). The other, Tc01 g036330, has reduplicated in cotton (Gorai.004G157600.1 and Gorai.001G169700.1). Asterisk indicates increased gene expression in elite versus wild tetraploids (Supplementary 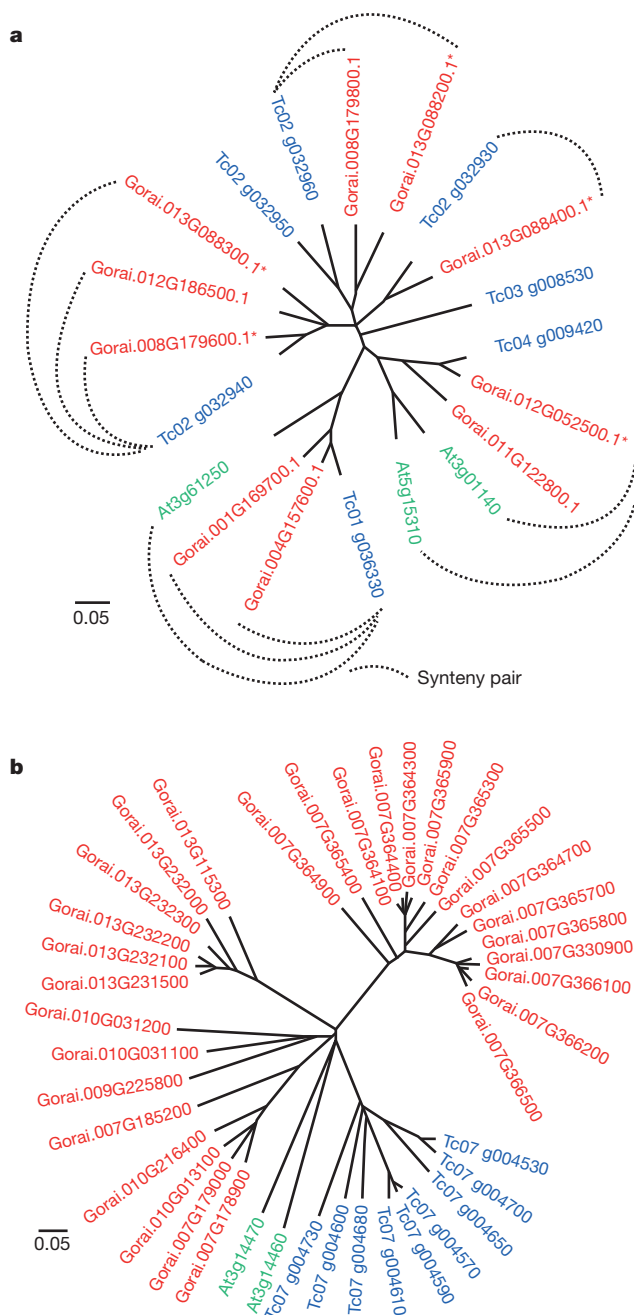Table 5.3). **b**, The most NBS-rich region of *T. cacao*, on chromosome 7, corresponds to regions of *G. raimondii* chromosome triplets 2/10/13 and 7/9/4. Cacao chromosome 7 NBSs form a single branch, indicating lineage-specific expansion. *G. raimondii* chromosome 7 and 13 NBSs form distinct branches, indicating cluster/tandem duplication (gene numbers also reflect physical proximity of genes to one another).

A total of 300 (0.8%) *G. raimondii* genes encode nucleotide-binding site (NBS) domains (Supplementary Table 3.6), largely of coiled-coil (CC)-NBS and CC-NBS-leucine rich repeat subgroups (165, 55%). Like cereals[14], after paleopolyploidy *G. raimondii* evolved clusters of new

NBS-encoding genes. The most NBS-rich (21%) region of *T. cacao*, on chromosome 7, corresponds to parts of *G. raimondii* chromosome triplets 2/10/13 and 7/9/4. In total, 27% and 25% of 294 mapped *G. raimondii* NBS genes are on these parts of chromosomes 7 and 9, often clustered in otherwise gene-poor surroundings (Supplementary Fig. 2.2). Most NBS clusters are species and chromosome specific (Fig. 3b and Supplementary Table 3.7), indicating rapid turnover and/or concerted evolution after cotton paleopolyploidy. In total, 230 (76.7%) NBS-encoding genes have experienced striking mutations (as detailed below) in the A genome since A–F divergence, reflecting an ongoing plant–pathogen 'arms race' (Supplementary Table 3.8).

Changes in gene expression during domestication have contributed to the deposition of >90% cellulose in cotton fibres, single-celled models for studying cell wall and cellulose biogenesis[15]. *G. raimondii* has at least 15 cellulose synthase (*CESA*) sequences required for cellulose synthesis[16] (Supplementary Table 3.3), with four single-gene *Arabidopsis* clades having three (*CESA3*, required in expanding primary walls) or two (*CESA4*, *CESA7* and *CESA8*, each required in the thickening of secondary walls) clade members in *G. raimondii*[16]. *G. raimondii* has at least 35 cellulose-synthase-like (*CSL*) genes required for synthesis of cell wall matrix polysaccharides that surround cellulose microfibrils[16] (Supplementary Table 3.4), including one family (*CSLJ*) absent in *Arabidopsis*[16]. Elite tetraploids have higher expression than wild cottons in 6 (40%) of 15 *CESA* genes and 12 (34%) of 35 *CSL* genes (Supplementary Table 5.3).

A total of 364 *G. raimondii* microRNA precursors from 28 conserved and 181 novel families (Supplementary Table 3.12), are predicted regulators of 859 genes enriched for molecule binding factors, catalytic enzymes, transporters and transcription factors (Supplementary Fig. 3.11, 12). Four conserved and 35 novel mRNAs were specifically expressed in *G. hirsutum* fibres, respectively targeting 53 and 318 genes, most with homology to proteins involved in fibre development (Supplementary Table 3.14, 15). Among 183,690 short interfering RNAs (siRNAs) found, 33,348 (18.15%) were on chromosome 13 (Supplementary Fig. 3.12), a vast enrichment. Small RNA[17–19] biogenesis proteins include 13 argonaute, 6 dicer-like (*DCL*) and 5 RNA-dependent RNA polymerase orthologues (Supplementary Table 3.16). *G. raimondii* seems to be the first eudicot with two *DCL3* genes and two genes encoding RNA polymerase IVa (Supplementary Table 3.16), perhaps relating to control of its abundant retrotransposons.

From unremarkable hairs found on all *Gossypium* seeds, 'spinnable' fibres (fibres with a ribbon-like structure that allows for spinning into yarn) evolved in the A genome after divergence from the B, E and F genomes ~5–10 Myr ago[4] (Fig. 1). To clarify the evolution of spinnable fibres, we sequenced the *G. herbaceum* A and *G. longicalyx* F genomes, which respectively differ from *G. raimondii* by 2,145,177 single-nucleotide variations (SNVs) and 477,309 indels, and 3,732,370 SNVs and 630,292 indels.

Specific genes are implicated in initial fibre evolution by both whole-gene and individual-nucleotide analyses. Across entire genes, 36 *G. herbaceum*–*G. raimondii* and 11 *G. herbaceum*–*G. longicalyx* orthologue pairs show evidence of diversifying selection ($\omega > 1$, $P < 0.05$) (Supplementary Table 4.1). A notable example, with *G. herbaceum*–*G. raimondii* $\omega > 9$, is Gorai.009G035800, a germin-like protein that is differentially expressed between normal and naked-seed cotton mutants during fibre expansion[20] and between wild and elite *G. barbadense* at 10 days post-anthesis (DPA; Supplementary Table 5.3).

Among 114,202 SNVs in 29,015 *G. herbaceum* genes after *G. herbaceum*–*G. longicalyx* A–F divergence (using D as outgroup, so F is the same as D, and A differs from both), we identified striking mutations including 1,090 non-synonymous mutations in 959 genes comprising the most severe 1% of functional impacts inferred using a modified entropy function[21]; 3,525 frameshift mutations (3,021 genes), 1,077 (987) premature stops, 527 (513) splice-site mutations, 102 (102) initiation alterations and 95 (94) extended reading frames (Supplementary Table 4.2, 3). These striking mutations have an average genomic

distribution (Supplementary Fig. 2.2) but are over-represented in genes coding for cell-wall-associated, kinase or nucleotide-binding proteins (Supplementary Table 4.5).

Striking mutations in the A-genome lineage are enriched ($P = 2.6 \times 10^{-18}$; Supplementary Discussion, section 4.4) within fibre-related quantitative trait locus (QTL) hotspots in A_tD_t tetraploid cottons[22], suggesting that post-allopolyploidy elaboration of fibre development[1] involved recursive changes in A_t and new changes in D_t genes. Striking A-genome mutations have orthologues in 1,051 D_t and 951 A_t fibre QTL hotspots. Likewise, sequencing of G. hirsutum cultivar Acala Maxxa revealed 495 striking mutations in 391 genes, with 83 (21.2%) in D_t fibre QTL hotspots and 73 (18.7%) in A_t hotspots (Supplementary Table 4.6).

QTL hotspots affecting multiple fibre traits[22] may reflect coordinated changes in expression of functionally diverse cotton genes. A total of 671 (1.79%) genes with >100 reads per million reads were differentially expressed in fibres from wild versus domesticated G. hirsutum (mostly at 10 DPA) and/or G. barbadense (mostly at 20 DPA) (Supplementary Table 5.3). Among 48 genes upregulated in domesticated G. hirsutum at 10 DPA, 20 (42%) are among 1,582 (4.2%) genes within QTL hotspot $D_t09.2$ (ref. 22) affecting length, uniformity, and short-fibre content, with 13 (27%) out of 677 (1.8%) genes in homoeologous hotspot A_t09 affecting fibre elongation and fineness. Out of 45 genes downregulated in domesticated G. barbadense at 20 DPA, 16 (35.6%) map to $D_t09.2$, and 8 (17.7%) to A_t09. In 79% of cultivated G. barbadense, this A_t region (which was then thought to be on chromosome 5, and is now known to be on chromosome 9) has been unconsciously introgressed by plant breeders with G. hirsutum DNA, suggesting an important contribution to productivity of G. barbadense cultivars[23].

A putative nuclear mitochondrial DNA (NUMT) sequence block[24] has an intriguing relationship with fibre improvement. A G. raimondii chromosome 1 region includes many genes closely resembling mitochondrial homologues ($K_s \sim 0.22$; Supplementary Table 4.7a). NUMT genes experienced a coordinated change in expression associated with G. barbadense domestication. The 105 (0.2%) genes upregulated in 10 DPA fibre of wild (versus elite) tetraploid G. barbadense (Supplementary Table 5.3) include 30 (37%; $P < 0.001$) of the 81 NUMT genes, including 8 NADH dehydrogenase and 4 cytochrome-c-related genes. All are within the QTL hotspot D_t01 that affects fibre fineness, length, and uniformity[22], suggesting a fibre-specific change in electron transfer in G. barbadense domestication.

Emergent features of polyploids may be related to processes that render them no longer the sum of their progenitors and permit them to explore transgressive phenotypic innovations. Despite the A-genome origin of spinnable fibres, after 1–2 Myr of co-habitation in tetraploid nuclei most A_t and D_t homoeologues are now expressed in fibres at similar levels (Supplementary Table 5.4). Such convergence is not ubiquitous: gene families involved in the synthesis of seed oil show strong A bias in wild G. hirsutum and its sister G. tomentosum, but strong D bias in an improved G. hirsutum (Supplementary Table 5.6).

Recruitment of D_t-genome genes into tetraploid fibre development[1] may have involved non-reciprocal DNA exchanges from A_t genes. In the ~40% of Acala Maxxa A_t and D_t genes that differ in sequence from their diploid progenitors (Fig. 4), most mutations are convergent, with A_t genes converted to the D_t state at more than twice the rate (25%) as the reciprocal (10.6%). Known to occur between cereal paralogues diverged by 70 Myr[14], non-reciprocal DNA exchanges are more abundant between cotton A_t and D_t genes separated by only ~5–10 Myr[4]. Such non-reciprocal exchanges explain prior observations including incongruent gene tree topology for 10% (3 pairs) of G. hirsutum A_t and D_t homoeologues in sequenced bacterial artificial chromosomes (BACs) (Supplementary Discussion, section 5.3); 13.2% of tetraploid DNA markers that showed different subgenomic affinities compared with the chromosomes to which they mapped, 9 of 13 being D_t biased (A_t to D_t)[25]; and expressed-sequence-tag-based
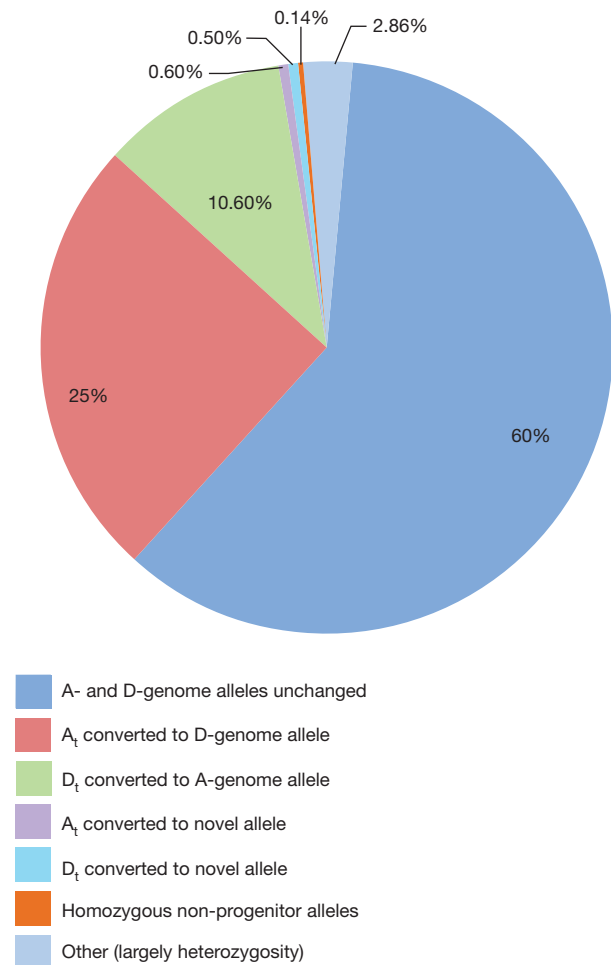

Figure 4 | Allelic changes between A- and D-genome diploid progenitors and the A_t and D_t subgenomes of G. hirsutum cultivar Acala Maxxa.

evidence of phylogenetic incongruity for as many as 7% of homeologous genes[26].

Several factors may have favoured D_t-biased allele conversion in tetraploid cotton. The nascent polyploid may have gained fitness from D-genome alleles native to its New World habitat. Before fortifying its reproductive barriers, the nascent polyploid may have occasionally outcrossed to nearby D-genome diploids, increasing the likelihood of illegitimate recombination. Outcrossing may also have contributed to the origin of Gossypium gossypioides, sister to G. raimondii and the only D-genome cotton containing many otherwise A-genome-specific repetitive DNAs[27–29]. D_t-biased allele conversion may have contributed to slightly greater protein-coding nucleotide diversity in the A_t compared with the D_t-genome (Supplementary Table 5.7).

Whereas the G. raimondii reference sequence and G. hirsutum short-read sequences reveal much about tetraploid cotton genome structure and polyploid evolution, high-contiguity sequencing of polyploids may elucidate still-cryptic features. Tetraploid cotton sequencing appears feasible: among six pairs of A_t and D_t BAC clones, the most similar pair shows 99.1% shared D_t-D and 97.6% A_t-D content (Supplementary Table 6.1), sufficient divergence to de-convolute shotgun sequence to the correct subgenome. Increased knowledge of molecular diversity is a foundation for integrating genomics with ecological and field-level knowledge of Gossypium species and their diverse adaptations to warm arid ecosystems on six continents.

METHODS SUMMARY

Sequencing. Reads were collected from Applied Biosystems 3730xl, Roche 454 XLR and Illumina Genome Analyzer IIx machines at the Joint Genome Institute

(http://www.jgi.doe.gov/sequencing/protocols/prots_production.html) or Hudson-Alpha Institute and Beckman Coulter Genomics (BAC end sequence), and USDA-ARS Mid-South Area Genomics Laboratory (*G. longicalyx*, *G. arboreum* and *G. hirsutum*).

Assembly. Assembly of 80,765,952 sequence reads used a modification of Arachne v.20071016, integrating linear (15× genome coverage) and paired (3.1× genome coverage) Roche 454 libraries corrected using 41.9 Gb Illumina sequence, with 1.54× paired-end Sanger sequences from two subclone, six fosmid and two BAC libraries. Cotton genetic and physical maps, and *Vitis vinifera* and *T. cacao* synteny were used to identify 51 joins across 64 scaffolds to form the 13 chromosomes (Supplementary Discussion, section 1). The remaining scaffolds were screened for contamination to produce a final assembly of 1,033 scaffolds (19,735 contigs) and 761.4 Mb. Sequences are in NCBI for *G. raimondii* (BioProject accession PRJNA171262), *G. longicalyx* (accession F1-1, SRA061660), *G. herbaceum* (accession A1-97, SRA061243) and *G. hirsutum* (cultivar Acala Maxxa, SRS375727) genomes; *G. hirsutum* (SRA061240) and *G. barbadense* (SRA061309) fibre transcriptomes; *G. hirsutum* (SRA061456) seed transcriptomes; and *G. hirsutum* microRNAs (SRA061415).

Annotation. PERTRAN software was used to construct transcript assemblies from ∼1.1 billion pairs of *G. raimondii* paired-end Illumina RNA-seq reads, 250 million *G. raimondii* single end reads, and 150 million *G. hirsutum* single end reads. PASA[30] was used to build transcript assemblies from 454 and Sanger resources (Supplementary Table 2.3). Loci were determined by transcript assembly and/or EXONERATE alignments of *A. thaliana*, cacao, rice, soybean, grape and poplar peptides to repeat-soft-masked *G. raimondii* genome using RepeatMasker. Gene models were predicted by three homology-based predictors (Supplementary Discussion, section 2.2). Best-scoring gene predictions were improved by PASA, then filtered on the basis of peptide homology or expressed-sequence-tag evidence to remove Pfam transposable element domain models. ClustalW alignments of amino acid sequences (Fig. 3) were used to guide coding sequence alignments. Phylogenetic trees were constructed by bootstrap neighbour-joining with a Kimura 2-parameter model using ClustalW2, assessing internal nodes with 1,000 replicates.

Full Methods and any associated references are available in the online version of the paper.

1. Jiang, C., Wright, R. J., El-Zik, K. M. & Paterson, A. H. Polyploid formation created unique avenues for response to selection in *Gossypium* (cotton). *Proc. Natl Acad. Sci. USA* **95,** 4419–4424 (1998).
2. Wendel, J. F. New world tetraploid cottons contain old-world cytoplasm. *Proc. Natl Acad. Sci. USA* **86,** 4132–4136 (1989).
3. Wang, X. et al. The genome of the mesopolyploid crop species *Brassica rapa*. *Nature Genet.* **43,** 1035–1139 (2011).
4. Senchina, D. S. et al. Rate variation among nuclear genes and the age of polyploidy in *Gossypium*. *Mol. Biol. Evol.* **20,** 633–643 (2003).
5. Wang, K. et al. The draft genome of a diploid cotton *Gossypium raimondii*. *Nature Genet.* **44,** 1098–1103 (2012).
6. Carvalho, M. R., Herrera, F. A., Jaramillo, C. A., Wing, S. L. & Callejas, R. Paleocene Malvaceae from northern South America and their biogeographical implications. *Am. J. Bot.* **98,** 1337–1355 (2011).
7. Jaillon, O. et al. The French–Italian Public Consortium for Grapevine Genome Characterization. The grapevine genome sequence suggests ancestral hexaploidization in the major angiosperm phyla. *Nature* **449,** 463–467 (2007).
8. Bowers, J. E., Chapman, B. A., Rong, J. & Paterson, A. H. Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. *Nature* **422,** 433–438 (2003).
9. Muravenko, O. V. et al. Comparison of chromosome BrdU–Hoechst–Giemsa banding patterns of the A1 and (AD)2 genomes of cotton. *Genome* **41,** 616–625 (1998).
10. Jiao, Y. et al. A genome triplication associated with early diversification of the core eudicots. *Genome Biol.* **13,** R3 (2012).
11. Fawcett, J. A., Maere, S. & Van de Peer, Y. Plants with double genomes might have had a better chance to survive the Cretaceous–Tertiary extinction event. *Proc. Natl Acad. Sci. USA* **106,** 5737–5742 (2009).
12. Stracke, R., Werber, M. & Weisshaar, B. The *R2R3-MYB* gene family in *Arabidopsis thaliana*. *Curr. Opin. Plant Biol.* **4,** 447–456 (2001).
13. Adkisson, P. L., Niles, G. A., Walker, J. K., Bird, L. S. & Scott, H. B. Controlling Cotton's insect pests: a new system. *Science* **216,** 19–22 (1982).
14. Wang, X., Tang, H. & Paterson, A. H. Seventy million years of concerted evolution of a homoeologous chromosome pair, in parallel in major Poaceae lineages. *Plant Cell* **23,** 27–37 (2011).
15. Haigler, C. H., Betancur, L., Stiff, M. R. & Tuttle, J. R. Cotton fiber: a powerful single-cell model for cell wall and cellulose research. *Front. Plant Sci.* **3,** 1–7 (2012).
16. Doblin, M. S., Pettolino, F. & Bacic, A. Plant cell walls: the skeleton of the plant world. *Funct. Plant Biol.* **37,** 357–381 (2010).
17. Baulcombe, D. RNA silencing in plants. *Nature* **431,** 356–363 (2004).
18. Matzke, M. A. & Birchler, J. A. RNAi-mediated pathways in the nucleus. *Nature Rev. Genet.* **6,** 24–35 (2005).
19. Brodersen, P. & Voinnet, O. The diversity of RNA silencing pathways in plants. *Trends Genet.* **22,** 268–280 (2006).
20. Kim, H. J. & Triplett, B. A. Cotton fiber germin-like protein. I. Molecular cloning and gene expression. *Planta* **218,** 516–524 (2004).
21. Reva, B., Antipin, Y. & Sander, C. Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic Acids Res.* **39,** e118 (2011).
22. Rong, J. et al. Meta-analysis of polyploid cotton QTL shows unequal contributions of subgenomes to a complex network of genes and gene clusters implicated in lint fiber development. *Genetics* **176,** 2577–2588 (2007).
23. Wang, G. L., Dong, J. M. & Paterson, A. H. The distribution of *Gossypium hirsutum* chromatin in *G. barbadense* germ plasm: molecular analysis of introgressive plant-breeding. *Theor. Appl. Genet.* **91,** 1153–1161 (1995).
24. Richly, E. & Leister, D. NUMTs in sequenced eukaryotic genomes. *Mol. Biol. Evol.* **21,** 1081–1084 (2004).
25. Reinisch, A. J. et al. A detailed RFLP map of cotton (*Gossypium hirsutum* × *Gossypium barbadense*): chromosome organization and evolution in a disomic polyploid genome. *Genetics* **138,** 829–847 (1994).
26. Flagel, L. E., Wendel, J. F. & Udall, J. A. Duplicate gene evolution, homoeologous recombination, and transcriptome characterization in allopolyploid cotton. *BMC Genomics* **13,** 302 (2012).
27. Wendel, J. F., Schnabel, A. & Seelanan, T. An unusual ribosomal DNA sequence from *Gossypium gossypioides* reveals ancient, cryptic, intergenomic introgression. *Mol. Phylogenet. Evol.* **4,** 298–313 (1995).
28. Zhao, X. P. et al. Dispersed repetitive DNA has spread to new genomes since polyploid formation in cotton. *Genome Res.* **8,** 479–492 (1998).
29. Cronn, R., Small, R. L., Haselkorn, T. & Wendel, J. F. Cryptic repeated genomic recombination during speciation in *Gossypium gossypioides*. *Evolution* **57,** 2475–2489 (2003).
30. Haas, B. J. et al. Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res.* **31,** 5654–5666 (2003).

Supplementary Information is available in the online version of the paper.

METHODS

Sequencing. Reads were collected with standard protocols (http://www.jgi.doe.gov/sequencing/protocols/prots_production.html) on Applied Biosystems 3730xl, Roche 454 XLR and Illumina Genome Analyzer (GA)IIx machines at the US Department of Energy Joint Genome Institute. Linear 454 data included standard XLR (47 runs, 16.868 Gb) and pre-release FLX+ data (5 runs, 3.262 Gb). Eight paired 454, 3–4-kilobase (kb) average insert size and one paired 12-kb average insert size were sequenced on standard XLR (23 runs, 5.931 Gb). One standard 400-base pair (bp) fragment library was sequenced at 2×150 (7 channels, 41.9 Gb) on an Illumina GAIIx. One 2.5-kb average insert size (405,024 reads, 286.1 Mb), one 6.5-kb average insert size library (374,125 reads, 263.0 Mb), six fosmid libraries (1,222,643 reads, 702.1 Mb) of 34–39-kb insert size, and two BAC libraries (107,520 reads, 77.5 Mb) of 98-kb and 115-kb (73,728 reads, 48.8 Mb) average insert size were sequenced on both ends for a total of 2,183,240 Sanger reads of 1.38 Gb of high-quality bases. FLX+ data were collected at the Roche Service Center. BAC end sequence (BES) was collected using standard protocols at the HudsonAlpha Institute.

Genome assembly and construction of pseudomolecule chromosomes. Organellar reads were removed by screening against mitochondria, chloroplast and ribosomal DNA. Any Roche 454 linear read <200 bp was discarded. Roche 454 paired reads in which either was shorter than 50 bp were discarded. An additional de-duplication step was applied to the 454 paired libraries that identifies and retains only one copy of each PCR duplicate. All remaining 454 reads were compared against a full Illumina GA2x run and any insertion/deletions in the 454 reads were corrected to match the Illumina alignments. The sequence reads were assembled using our modified version of Arachne v.20071016 (ref. 31) with parameters maxcliq1 = 100, correct1_passes = 0 and BINGE_AND_PURGE = True, bless = False maxcliq1 = 200 BINGE_AND_PURGE = True lap_ratio = 0.8 max_bad_look = 1000 (note Arachne error correction is on). This produced 1,263 scaffold sequences, with a scaffold L50 of 25.8 Mb, 58 scaffolds larger than 100 kb, and total genome size of 761.8 Mb. Scaffolds were screened against bacterial proteins, organelle sequences and non-redundant GenBank and removed if found to be a contaminant. Additional scaffolds were removed if they: (1) consisted of >95% 24-nucleotide sequences that occurred four other times in scaffolds larger than 50 kb; (2) contained only unanchored RNA sequences; or (3) were <1 kb in length.

The combination of BES/markers hybridized to fingerprint contigs[32], 2,800 markers in a genetic map for the D genome in an A_tD_t plant[33] and 262 markers from the tetraploid genetic map[34], along with *Vitis vinifera* and *T. cacao* synteny was used to identify breaks in the initial assembly. Markers were aligned to the assembly using BLAT[35] (parameters: -t = dna -q = dna −minScore = 200 −extendThroughN). BES, physical map contigs, *V. vinifera* and *T. cacao* genes were aligned to the genome using BLAST[36]. Scaffolds were broken if they contained linkage group/syntenic discontinuity coincident with an area of low BAC/fosmid coverage. A total of 13 breaks were executed, and 64 of the broken scaffolds were oriented, ordered and joined using 51 joins to form the final assembly containing 13 pseudomolecule chromosomes. Each chromosome join is padded with 10,000 missing nucleotides. The final assembly contains 1,033 scaffolds (19,735 contigs) that cover 761.4 Mb of the genome with a contig L50 of 135.6 kb and a scaffold L50 of 62.2 Mb.

The assembly size is near the centre of genome-size estimates of 880 Mb from flow cytometry[37], 630 Mb from Feulgen cytophotometry[38], and 650 Mb[39] and 770 Mb[40] from re-naturation kinetics.

Completeness of the euchromatic portion of the genome assembly was assessed using 65,506 *G. raimondii* complementary DNAs obtained from GenBank, which were aligned to the assembly using BLAT[3] (parameters: -t = dna -q = rna −extendThroughN). The aim of the completeness analysis was to obtain a measure of completeness of the assembly, rather than a comprehensive examination of gene space. cDNAs were aligned to the assembly using BLAT[35] (parameters: -t = dna -q = rna −extendThroughN) and alignments that comprised ≥90% base-pair identity and ≥85% EST coverage were retained. The screened alignments indicate that 57,170 out of 63,506 (90.3%) cDNAs aligned to the assembly. The cDNAs that failed to align were primarily composed of stretches of polynucleotide sequences that failed to generate non-random alignments to any plant or other organism in the NCBI as of the release date.

Annotation. A total of 85,746 transcript assemblies were constructed from about 1.1 billion pairs of D5 paired-end Illumina RNA-seq reads, 55,294 transcript assemblies from 250 million D5 single-end Illumina RNA-seq reads and 62,526 transcript assemblies from 150 million *G. hirsutum* cotton single-end Illumina RNA-seq reads. All these transcript assemblies were constructed using PERTRAN software (in preparation). In total, 120,929 transcript assemblies were built using PASA[30] from 56,638 D5 Sanger ESTs, 2.5 million D5 Roche 454 RNA-seq reads and all of the RNA-seq transcript assemblies. An additional 133,073 transcript assemblies were constructed using PASA from 296,214 *G. hirsutum* cotton Sanger ESTs and about 2.9 million *G. hirsutum* cotton 454 reads. The larger number of transcript assemblies from fewer *G. hirsutum* sequences is due to the fragmented nature of the assemblies. Loci were determined by transcript assembly alignments and/or EXONERATE alignments of peptides from *A. thaliana*, cacao, rice, soybean, grape and poplar peptides to repeat-soft-masked D5 genome using RepeatMasker (http://www.repeatmasker.org) with up to 2,000-bp extensions on both ends, unless extending into another locus on the same strand. Gene models were predicted by homology-based predictors, FGENESH+[41], FGENESH_EST (similar to FGENESH+, EST as splice site and intron input instead of peptide/translated open-reading frames) and GenomeScan[42]. The best scored predictions for each locus are selected using multiple positive factors including EST and peptide support, and one negative factor: overlap with repeats. The selected gene predictions were improved by PASA. Improvement includes adding untranslated regions, splicing correction, and adding alternative transcripts. PASA-improved gene model peptides were subject to peptide-homology analysis to above-mentioned proteomes in order to obtain Cscore and peptide coverage. Cscore is a peptide BLASTP score ratio mutual best hit BLASTP score and peptide coverage is highest percentage of peptide aligned to the best of homologues. PASA-improved transcripts were selected on the basis of Cscore, peptide coverage, EST coverage and its coding sequence (CDS) overlapping with repeats. The transcripts were selected if their Cscore was larger than or equal to 0.5 and peptide coverage larger than or equal to 0.5, or if it had EST coverage, but its CDS overlapping with repeats was less than 20%. For gene models whose CDS overlaps with repeats for more than 20%, its Cscore needed to be at least 0.9 and homology coverage at least 70% to be selected. The selected gene models were subject to Pfam analysis and gene models whose peptide was more than 30% in Pfam transposable element domains were removed. The final gene set had 37,505 protein-coding genes and 77,267 protein-coding transcripts.

31. Jaffe, D. B. *et al.* Whole-genome sequence assembly for mammalian genomes: Arachne 2. *Genome Res.* **13,** 91–96 (2003).
32. Lin, L. *et al.* A draft physical map of a D-genome cotton species (*Gossypium raimondii*). *BMC Genomics* **11,** 389–417 (2010).
33. Rong, J.-K. *et al.* A 3347-locus genetic recombination map of sequence-tagged sites reveals features of genome organization, transmission and evolution of cotton (*Gossypium*). *Genetics* **166,** 389–417 (2004).
34. Rong, J. *et al.* Comparative genomics of *Gossypium* and *Arabidopsis*: unraveling the consequences of both ancient and recent polyploidy. *Genome Res.* **15,** 1198–1210 (2005).
35. Kent, W. J. BLAT–the BLAST-like alignment tool. *Genome Res.* **4,** 656–664 (2002).
36. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215,** 403–410 (1990).
37. Hendrix, B. & Stewart, J. M. Estimation of the nuclear DNA content of *Gossypium* species. *Ann. Bot.* **95,** 789–797 (2005).
38. Kadir, Z. B. A. DNA evolution in the genus *Gossypium*. *Chromosoma* **56,** 85–94 (1976).
39. Geever, R., Katterman, F. & Endrizzi, J. DNA hybridization analyses of *Gossypium* allotetraploid and two closely related diploid species. *Theor. Appl. Genet.* **77,** 553–559 (1989).
40. Walbot, V. & Dure, L. S. Developmental biochemistry of cotton seed embryogenesis and germination. VII. Characterization of cotton genome. *J. Mol. Biol.* **101,** 503–536 (1976).
41. Salamov, A. A. & Solovyev, V. V. *Ab initio* gene finding in *Drosophila* genomic DNA. *Genome Res.* **10,** 516–522 (2000).
42. Yeh, R.-F., Lim, L. P. & Burge, C. Computational inference of homologous gene structures in the human genome. *Genome Res.* **11,** 803–816 (2001).